Morphological filters for OCR: a performance comparison

Laurent Mennillo[1], Jean Cousty[2] and Laurent Najman[2]

[1]Institut Pascal - UMR6602 - UBP/CNRS/IFMA
[2]Université Paris-Est, LIGM, Équipe A3SI, ESIEE, France

Abstract

In this article is compared the ability of several morphological operators to improve OCR performance when used as preprocessing filters. An experiment on binary and greyscale images using the Tesseract OCR engine and morphological filters acting in complex, graph and vertex spaces has thus been performed and results in a good overall performance of complex and area filters. MSE measures have also been performed to evaluate the denoising ability of these filters, which again shows the good performance of both complex and area filters on this aspect.

1 Introduction

This article presents an OCR performance comparison obtained through the use of several morphological filters on degraded documents. First, section 2 will present the degradation models used to alter the documents to be analysed, section 3 will then introduce morphological filters partially restoring the image quality of such degraded documents, section 4 will describe the detailed test protocols of this experiment and the last sections will finally discuss the results and conclude.

2 Document degradation models

Document degradation models are designed to simulate local distortions that are introduced during the processes of document scanning, printing and photocopying. In the context of this experiment, some of these models have the ability to generate realistic degradations that are appropriate for OCR performance evaluation. Besides, increasing levels of such degradations can also be produced by adjusting the models parameters, thus allowing a proper leveled comparison.

2.1 Binary document degradation

The binary document degradation model used in this experiment has been described by Kanungo *et al.* in [3]. This local model, which only applies to binary images, accounts for two types of document degradation, which are *pixel inversion* and *blurring*. *Pixel inversion* simulates image noise usually generated by light intensity variations, sensor sensitivity and image thresholding, while *blurring* simulates the point-spread function of the scanner optical system. *Pixel inversion* probability of a background (*resp.* foreground) pixel is modelled following an exponential function of its distance from the nearest foreground (*resp.* background) pixel as:

$$p(0|1,d,\alpha_0,\alpha) = \alpha_0 e^{-\alpha d^2} + \eta \qquad p(1|0,d,\beta_0,\beta) = \beta_0 e^{-\beta d^2} + \eta$$

Parameter d is thus the 4-neighbor distance of each background (*resp.* foreground) pixel from its nearest foreground (*resp.* background) pixel, parameter α_0 (*resp.* β_0) is the amount of generated noise related to background (*resp.* foreground) pixels, parameter α (*resp.* β) is the decay speed, relatively to distance d, of background (*resp.* foreground) pixels flipping probability and parameter η is the constant flipping probability of all pixels. Finally, *Blurring* is produced by a morphological closing operation using a disk structuring element of diameter k. The described model with parameters $\Theta = (\eta, \alpha_0, \alpha, \beta_0, \beta, k)$ is thus used to degrade any binary document by computing the distance map of each pixel, then independently flip them following their respective probability and finally perform a morphological closing operation.

2.2 Greyscale document degradation

The document degradation model described in 2.1 can also be used to process greyscale images. For this purpose, the image to degrade is first thresholded with value $t \in [0; 255]$ to binary images B_t. These images are then degraded with the binary document degradation model. Thereafter, greyscale images G_t are generated by setting the black pixels of images B_t to their respective tresholding value t. The degraded image D, initialized with white pixels, is finally reconstructed from white to black values ($t = 255 \ldots 0$) by processing the minimum value at each pixel of images D and G_t ($D_{i,j} = min(D_{i,j}, G_{t,i,j})$). In this experiment, a faster but less realistic greyscale document degradation is achieved by adding impulsional Gaussian noise to the original image with parameters $\Omega = (\sigma, p)$, where σ and p, respectively denote the standard deviation and proportion of pixels to alter.

3 Morphological filtering

Morphological filters can be used to restore or improve the image quality of digitally converted documents and thus increase OCR performance. This section presents the four morphological filters compared in this experiment.

3.1 Morphological operators in simplicial complex spaces

Morphological operators defined in simplicial complex spaces have been presented by Dias *et al.* in [2]. The following paragraphs will introduce these morphological operators and describe their implementation.

3.1.1 Definitions

Lattice A *lattice* is a partially ordered set that have a unique lowest greater bound called *supremum* and a greatest lower bound called *infimum*. An example of lattice, given in figure 1, is the power set $\mathcal{P}(X)$ of set $X = \{a, b, c\}$, whose supremum and infimum are respectively the union and the intersection.

Adjunction Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be two lattices whose order relations are denoted by $\leq_1$ and $\leq_2$. Two operators $\alpha : \mathcal{L}_2 \to \mathcal{L}_1$ and $\alpha^A : \mathcal{L}_1 \to \mathcal{L}_2$ form an *adjunction* (α^A, α) if, $\forall a \in \mathcal{L}_2$ and $\forall b \in \mathcal{L}_1$, $\alpha(a) \leq_1 \Leftrightarrow\ a \leq_2 \alpha^A(b)$.

Simplex An *(abstract) simplex* of dimension $n \in \mathbb{N}$, or n-simplex, is a finite nonempty set. The dimension of a simplex x is denoted by $dim(x)$. Geometrically, 0-simplices can be represented as

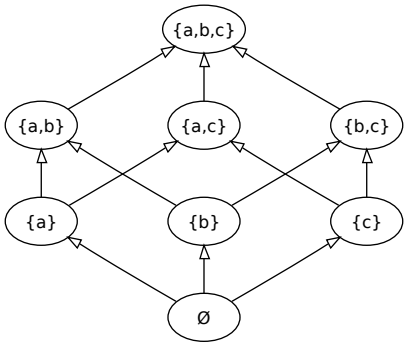
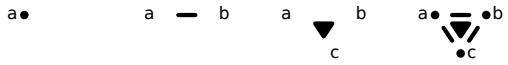


Figure 1: Power set $\mathcal{P}(X)$ of set $X = \{a, b, c\}$.

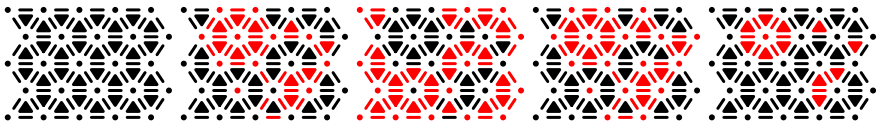
points, 1-simplices as segments, 2-simplices as triangles and so forth. Illustrations of simplices $\{a\}$, $\{a, b\}$ and $\{a, b, c\}$ are respectively given in figures 2a, 2b and 2c.



(a) 0-simplex (b) 1-simplex (c) 2-simplex (d) 2-cell

Figure 2: Graphical representation of simplices and cell.

Simplicial complex An *(abstract) simplicial complex*, or complex, is a set X of simplices such that, for any set $x \in X$, any nonempty subset of x also belongs to X. Its dimension $n \in \mathbb{N}$ is equal to the greatest dimension of its simplices. In the following, $\mathbb{C}$ denotes a nonempty complex of dimension n, or n-complex. The *complement* of a subset X of $\mathbb{C}$ is $\overline{X} = \mathbb{C} \setminus X$, while the set of all subsets of $\mathbb{C}$, equipped with the inclusion relation and denoted by $\mathcal{P}(\mathbb{C})$, is a lattice, which is complemented. Figures 3a, 3b and 3c respectively presents a complex of dimension 2, or 2-complex, a subset X and its complement $\overline{X}$.



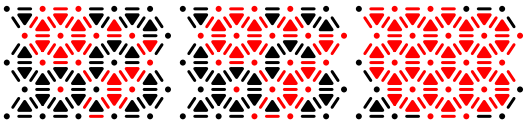
(a) $\mathbb{C}$ (b) X (c) $\overline{X}$ (d) $\mathcal{C}$ (e) $\mathcal{S}$

Figure 3: Complex, subset, complement, subcomplex and star.

Subcomplex, star and cell Any subset X of $\mathbb{C}$ that is also a complex is called a *subcomplex*. A *star* is a subset X of $\mathbb{C}$ whose complement $\overline{X}$ is a subcomplex. The set of all subcomplexes and stars of $\mathbb{C}$ are respectively denoted by $\mathcal{C}$ and $\mathcal{S}$. Note that the sets $\mathcal{C}$ and $\mathcal{S}$, equipped with the inclusion relation, are sublattices of $\mathcal{P}(\mathbb{C})$ that are closed under union and intersection, but not complemented. A subcomplex X of $\mathbb{C}$ is called a *cell* if there exists a simplex $x \in X$ such that X is exactly the set of all subsets of x. Following this definition, cells can be considered as

the elementary building blocks of simplicial complexes. Illustrations of a subcomplex, a star and a cell are respectively given in figures 3d, 3e and 2d.

Closure and star of a set The *closure* of a set X of $\mathbb{C}$, denoted $Cl(X)$, is the smallest subcomplex of $\mathbb{C}$ that contains each simplex in X. The *star* of a set X of $\mathbb{C}$, denoted $St(X)$, is the set of all simplices in $\mathbb{C}$ that intersect with X. Figures 4a, 4b and 4c present respectively a set X, its closure $Cl(X)$ and its star $St(X)$.



(a) X (b) $Cl(X)$ (c) $St(X)$

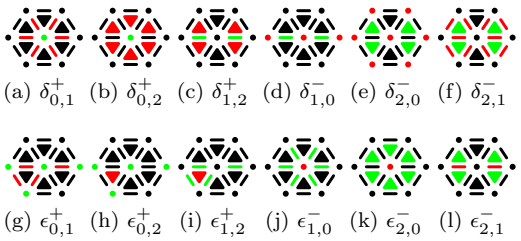
Figure 4: Closure and star of a set.

3.1.2 Dimensional operators

Dimensional operators are designed to add or remove simplices from a set by taking into account their dimension. Let $X \in \mathbb{C}$ and let $i \in [0; n]$, n being the dimension of $\mathbb{C}$. The set of all i-simplices of X is $X_i = \{x \in X \mid dim(x) = i\}$. In particular, $\mathbb{C}_i$ is the set of all i-simplices of $\mathbb{C}$ and $\mathcal{P}(\mathbb{C}_i)$ is the set of all subsets of $\mathbb{C}_i$, $\mathcal{P}(\mathbb{C}_i)$ being complemented. Now, let $i, j \in \mathbb{N}$ such that $i \leq j \leq n$. The operators $\delta^+_{i,j}$ and $\epsilon^+_{i,j}$ act from $\mathcal{P}(\mathbb{C}_i)$ into $\mathcal{P}(\mathbb{C}_j)$ and the operators $\delta^-_{j,i}$ and $\epsilon^-_{j,i}$ act from $\mathcal{P}(\mathbb{C}_j)$ into $\mathcal{P}(\mathbb{C}_i)$ as:

$\mathcal{P}(\mathbb{C}_i) \to \mathcal{P}(\mathbb{C}_j)$	$\mathcal{P}(\mathbb{C}_j) \to \mathcal{P}(\mathbb{C}_i)$
$X \to \delta^+_{i,j}(X)$ *such that* $\delta^+_{i,j}(X) = \{x \in \mathbb{C}_j \mid \exists y \in X, y \subseteq x\}$	$X \to \delta^-_{j,i}(X)$ *such that* $\delta^-_{j,i}(X) = \{x \in \mathbb{C}_i \mid \exists y \in X, x \subseteq y\}$
$X \to \epsilon^+_{i,j}(X)$ *such that* $\epsilon^+_{i,j}(X) = \{x \in \mathbb{C}_j \mid \forall y \in \mathbb{C}_i, y \subseteq x \Rightarrow y \in X\}$	$X \to \epsilon^-_{j,i}(X)$ *such that* $\epsilon^-_{j,i}(X) = \{x \in \mathbb{C}_i \mid \forall y \in \mathbb{C}_j, x \subseteq y \Rightarrow y \in X\}$

Note that the pair of operators $(\epsilon^-_{j,i}, \delta^+_{i,j})$ and $(\epsilon^+_{i,j}, \delta^-_{j,i})$ are adjunctions acting between $\mathcal{P}(\mathbb{C}_i)$ and $\mathcal{P}(\mathbb{C}_j)$ and that operators $\delta^+_{i,j}$ and $\epsilon^+_{i,j}$ (*resp.* $\delta^-_{j,i}$ and $\epsilon^-_{j,i}$) are dual w.r.t. the complement. These dimensional operators are illustrated in figure 5, where green simplices are the original set and red simplices are the resulting set.



(a) $\delta^+_{0,1}$ (b) $\delta^+_{0,2}$ (c) $\delta^+_{1,2}$ (d) $\delta^-_{1,0}$ (e) $\delta^-_{2,0}$ (f) $\delta^-_{2,1}$

(g) $\epsilon^+_{0,1}$ (h) $\epsilon^+_{0,2}$ (i) $\epsilon^+_{1,2}$ (j) $\epsilon^-_{1,0}$ (k) $\epsilon^-_{2,0}$ (l) $\epsilon^-_{2,1}$

Figure 5: Dimensional operators.

3.1.3 Closure and star operators

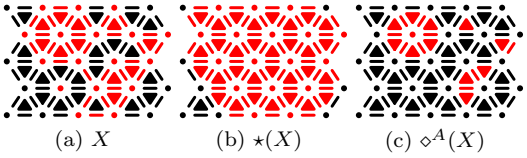
Dimensional operators previously described allow for the creation of $\diamond : \mathcal{S} \to \mathcal{C}$ and $\star : \mathcal{C} \to \mathcal{S}$ operators, defined by:

$$\forall X \in \mathcal{S},\ \diamond(X) = \bigcup_{i\in[0;n],j\in[0;n]} \delta^{-}_{j,i}(X) \qquad \forall X \in \mathcal{C},\ \star(X) = \bigcup_{i\in[0;n],j\in[0;n]} \delta^{+}_{i,j}(X)$$

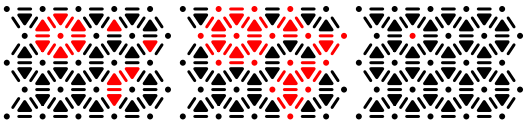
Note that $\diamond$ (*resp.* $\star$) operator, acting as Cl (*resp.* St) on $\mathcal{S}$ (*resp.* $\mathcal{C}$), produce subcomplexes (*resp.* stars). In words of mathematical morphology, these operators are *dilations*, meaning that they commute with the supremum. Their adjunct *erosions*, $\diamond^A : \mathcal{C} \to \mathcal{S}$ and $\star^A : \mathcal{S} \to \mathcal{C}$ operators, commuting with the infimum, are dual w.r.t. the complement and thus defined by:

$$\forall X \in \mathcal{C},\ \diamond^A(X) = \overline{\diamond(\overline{X})} \qquad \forall X \in \mathcal{S},\ \star^A(X) = \overline{\star(\overline{X})}$$

Note that $\diamond^A$ (*resp.* $\star^A$) operators, acting on $\mathcal{C}$ (*resp.* $\mathcal{S}$), produce stars (*resp.* subcomplexes). These operators, acting on the subcomplex X and the star Y, are illustrated in figure 6.



(a) X (b) $\star(X)$ (c) $\diamond^A(X)$



(d) Y (e) $\diamond(Y)$ (f) $\star^A(Y)$

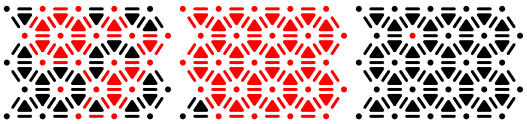
Figure 6: Closure and star operators.

3.1.4 Dilation and erosion operators

Composition of closure and star operators lead to $\delta : \mathcal{C} \to \mathcal{C}$ and $\epsilon : \mathcal{C} \to \mathcal{C}$ operators, defined by:

$$\forall X \in \mathcal{C},\ \delta(X) = \diamond \circ \star(X) \qquad \forall X \in \mathcal{C},\ \epsilon(X) = \star^A \circ \diamond^A(X)$$

Note that δ is a dilation, ϵ an erosion and that the pair (ϵ, δ) is an adjunction. δ and ϵ operators acting on the subcomplex X shown in figure 7a are respectively illustrated in figures 7b and 7c.



(a) X (b) $\delta(X)$ (c) $\epsilon(X)$

Figure 7: Dilation and erosion operators in complexes.

3.1.5 Closing and opening filters

A filter is an operator $\beta : \mathcal{L} \to \mathcal{L}$ that is increasing: $X \subseteq Y \Rightarrow \beta(X) \subseteq \beta(Y)$ and idempotent: $\beta \circ \beta(X) = \beta(X)$. Composition of dilation and erosion operators lead to $\phi : \mathcal{C} \to \mathcal{C}$ and $\gamma : \mathcal{C} \to \mathcal{C}$ filters, defined by:

$$\forall X \in \mathcal{C},\ \phi(X) = \epsilon \circ \delta(X) \qquad \forall X \in \mathcal{C},\ \gamma(X) = \delta \circ \epsilon(X)$$

In words of mathematical morphology, ϕ filter is a *closing* operation, which is extensive: $X \subseteq \phi(X)$, while γ filter is an *opening* operation, which is anti-extensive: $\gamma(X) \subseteq X$. These two filters are illustrated in figure 8.

(a) X (b) $\phi(X)$ (c) $\gamma(X)$

Figure 8: Closing and opening filters in complexes.

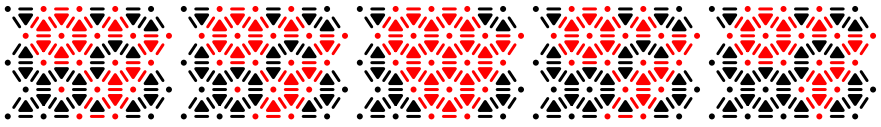
3.1.6 Dimensional-closing and dimensional-opening filters

Dimensional filters can also be created with dimensional operators. Given any dimension $d \in [0; n]$, $\phi_{d/(n+1)} : \mathcal{C} \to \mathcal{C}$ and $\gamma_{d/(n+1)} : \mathcal{C} \to \mathcal{C}$ filters are thus defined by:

$$\forall X \in \mathcal{C},\ \phi_{d/(n+1)}(X) = \bigcup \left[\{X_i \mid i \in [0; n-d-1]\} \cup \{\epsilon^+_{n-d,j}(X_d) \mid j \in [n-d; n]\} \right]$$

$$\forall X \in \mathcal{C},\ \gamma_{d/(n+1)}(X) = \bigcup \{\delta^-_{j,i}(X_j) \mid j \in [d; n], i \in [0; j]\}$$

In other words, $\phi_{d/(n+1)}$ is a closing operation which adds to the set X the cells of $\mathbb{C}$ whose i-simplices, $i \in [0; n-d]$, belong to X, while $\gamma_{d/(n+1)}$ is an opening operation which removes from the set X the cells of $\mathbb{C}$ whose dimension is less than d. Besides, note that $\phi_{d/(n+1)}$ and $\gamma_{d/(n+1)}$ filters satisfy the granulometric properties $\phi_{d_1/(n+1)}(X) \subseteq \phi_{d_2/(n+1)}(X)$ and $\gamma_{d_2/(n+1)}(X) \subseteq \gamma_{d_1/(n+1)}(X)$, if $d_1 \leq d_2$, $\forall d_1, d_2 \in [0; n-1]$. These filters are illustrated in figure 9.



(a) X (b) $\phi_{1/3}(X)$ (c) $\phi_{2/3}(X)$ (d) $\gamma_{1/3}(X)$ (e) $\gamma_{2/3}(X)$

Figure 9: Dimensional-closing and dimensional-opening filters.

3.1.7 Granulometries

Closing and opening filters can be extended to granulometries by composing them with dimensional filters. Let α be an operator defined on a lattice $\mathcal{L}$ and let i be a nonnegative integer. The operator $\alpha^i(\mathcal{L})$ is defined by the identity when $i = 0$ and by $(\alpha \circ \alpha^{i-1})(\mathcal{L})$ otherwise. Now,

let parameters i and d respectively be the quotient and the remainder of the division $k/(n+1)$. Granulometric-closing filter $\Phi_{k/(n+1)} : \mathcal{C} \to \mathcal{C}$ and granulometric-opening filter $\Gamma_{k/(n+1)} : \mathcal{C} \to \mathcal{C}$ are then defined by:

$$\forall X \in \mathcal{C},\ \Phi_{k/(n+1)} = \epsilon^i \circ \phi_{d/(n+1)} \circ \delta^i(X) \qquad \forall X \in \mathcal{C},\ \Gamma_{k/(n+1)} = \delta^i \circ \gamma_{d/(n+1)} \circ \epsilon^i(X)$$

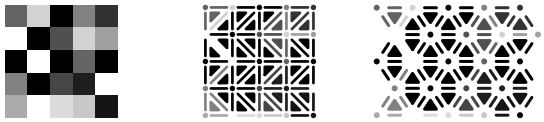
3.1.8 Alternating sequential filter

The alternating sequential filter $ASF_{k/(n+1)} : \mathcal{C} \to \mathcal{C}$ can be defined using the granulometric closing and opening filters by:

$$\forall X \in \mathcal{C},\ ASF_{k/(n+1)}(X) = \begin{cases} \text{identity} & \text{if } k = 0 \\ \Gamma_{k/(n+1)} \circ \Phi_{k/(n+1)} \circ ASF_{k-1/(n+1)}(X) & \text{otherwise} \end{cases}$$

3.1.9 Implementation

The simplicial complex space used in this experiment is the hexagonal grid $\mathbb{C}$ shown in figure 3a and initialized in figure 10c, on which each vertex is 6-connected. In order to work on square grids illustrated in figure 10a, which is the case of two-dimensional images, an other structure merging these two representations, as shown in figure 10b, has been used for complex filtering. Additionally, working with greyscale values is done by respectively computing the maximum and minimum value for union and intersection, while the complement of a simplex x is defined by the maximum value of the greyscale space minus the value of x. As for initialization, the set X_0 is initialized with the image pixels values, any 1-simplex $x \in X_1$ is set to the minimum value of its corresponding 0-simplices $x_1, x_2 \in \mathcal{P}(\mathbb{C}_0) \mid x_1, x_2 \subset x$ with $x_1 \neq x_2$, and any 2-simplex $y \in X_2$ is also set to the minimum value of its corresponding 0-simplices $y_1, y_2, y_3 \in \mathcal{P}(\mathbb{C}_0) \mid y_1, y_2, y_3 \subset y$ with $y_1 \neq y_2 \neq y_3$. Following this initialization, the initial set is thus always a subcomplex, which is coherent given the filters used in this experiment, all acting on set $\mathcal{C}$. One can note, however, that in the case of binary images, such initialized sets are invariant to dimensional-closing filter $\phi_{d/(n+1)}$. This particular property means that $ASF_{k/(n+1)}$ filter applied on such sets will not produce the expected result in this particular case, as the outcome of this operation will appear to be similar to the closing of an opening rather than the opening of a closing. Figure 10 illustrates the used data structure and its initialization.



(a) Image (b) Structure (c) Complex

Figure 10: Used data srtucture and initialization in complexes.

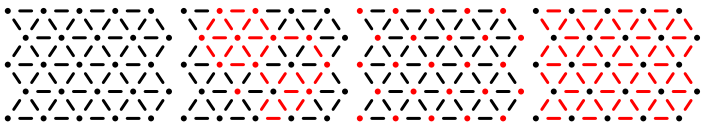
3.2 Morphological operators in graph spaces

Morphological operators defined in graph spaces have been presented by Cousty *et al.* in [1]. These operators being very similar to those used in simplicial complex spaces, a large part of the following is thus based on the theory already covered.

3.2.1 Definitions

Graph A *graph* $\mathbb{G}$ is a complex of dimension 1. 0-simplices of a graph are usually denoted as its vertex set, while 1-simplices are usually denoted as its edge set. Such graph is presented in figure 11a along with a subset, also denoted as a subgraph, in figure 11b.

Important notations In the following, the set of all subsets of $\mathbb{G}$, or $\mathcal{P}(\mathbb{G})$, is denoted by $\mathcal{G}$. Similarly, the set of all vertices (*resp.* edges) of $\mathbb{G}$, or $\mathcal{P}(\mathbb{G}_0)$ (*resp.* $\mathcal{P}(\mathbb{G}_1)$), is denoted by $\mathcal{G}_0$ (*resp.* $\mathcal{G}_1$). The sets $\mathcal{G}_0$ and $\mathcal{G}_1$ are respectively illustrated in figures 11c and 11d.



(a) $\mathbb{G}$ (b) X (c) $\mathcal{G}_0$ (d) $\mathcal{G}_1$

Figure 11: Graph, subgraph, vertices and edges.

3.2.2 Dimensional operators

As in simplicial complex spaces, dimensional operators in graph spaces are designed to add or remove vertices and edges from a set. These operators in graphs are exactly the same operators as those defined in complexes, excepted that their dimension d is in the interval $[0; 1]$.

3.2.3 Dimensional dilation and erosion operators

In graphs, one is often interested by operators whose input and output are a set of vertices (*resp.* edges). Composing the dimensional operators defined above lead to $\delta : \mathcal{G}_0 \to \mathcal{G}_0$ and $\epsilon : \mathcal{G}_0 \to \mathcal{G}_0$ operators, acting on vertices and defined by:

$$\forall X \in \mathcal{G}_0,\ \delta(X) = \delta^{-}_{1,0} \circ \delta^{+}_{0,1}(X) \qquad \forall X \in \mathcal{G}_0,\ \epsilon(X) = \epsilon^{-}_{1,0} \circ \epsilon^{+}_{0,1}(X)$$

Similarly to δ and ϵ operators, $\Delta : \mathcal{G}_1 \to \mathcal{G}_1$ and $\varepsilon : \mathcal{G}_1 \to \mathcal{G}_1$ operators, acting on edges, are defined by:

$$\forall X \in \mathcal{G}_1,\ \Delta(X) = \delta^{+}_{0,1} \circ \delta^{-}_{1,0}(X) \qquad \forall X \in \mathcal{G}_1,\ \varepsilon(X) = \epsilon^{+}_{0,1} \circ \epsilon^{-}_{1,0}(X)$$

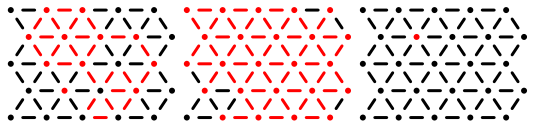
Note that δ and Δ operators are dilations, while ϵ and ε are erosions.

3.2.4 Dilation and erosion operators

Composing dimensional dilation and erosion operators lead to graph dilation and erosion operators. $(\delta \vee \Delta) : \mathcal{G} \to \mathcal{G}$ and $(\epsilon \vee \varepsilon) : \mathcal{G} \to \mathcal{G}$ operators are defined by:

$$\forall X \in \mathcal{G},\ (\delta \vee \Delta)(X) = \delta(X_0) \cup \Delta(X_1) \qquad \forall X \in \mathcal{G},\ (\epsilon \vee \varepsilon)(X) = \epsilon(X_0) \cup \varepsilon(X_1)$$

Note that $(\delta \vee \Delta)$ is a dilation, while $(\epsilon \vee \varepsilon)$ is an erosion. These operators are illustrated in figure 12.



(a) X (b) $(\delta \vee \Delta)(X)$ (c) $(\epsilon \vee \varepsilon)(X)$

Figure 12: Graph dilation and erosion operators.

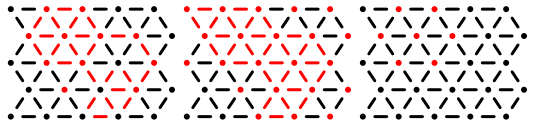
3.2.5 Closing and opening filters

Composition of δ, Δ, ϵ and ε operators lead to $(\phi \vee \Phi)_1 : \mathcal{G} \to \mathcal{G}$ and $(\gamma \vee \Gamma)_1 : \mathcal{G} \to \mathcal{G}$ filters, defined by:

$$\forall X \in \mathcal{G},\ (\phi \vee \Phi)_1(X) = (\epsilon \vee \varepsilon) \circ (\delta \vee \Delta)(X)$$

$$\forall X \in \mathcal{G},\ (\gamma \vee \Gamma)_1(X) = (\delta \vee \Delta) \circ (\epsilon \vee \varepsilon)(X)$$

Note that $(\phi \vee \Phi)_1$ is a closing operation, while $(\gamma \vee \Gamma)_1$ is an opening operation. These filters are illustrated in figure 13.



(a) X (b) $(\phi \vee \Phi)_1(X)$ (c) $(\gamma \vee \Gamma)_1(X)$

Figure 13: Graph closing and opening filters.

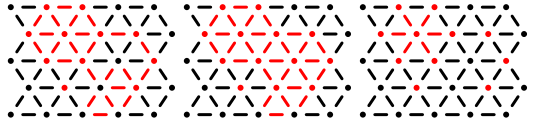
3.2.6 Half-closing and half-opening filters

Half opening and closing filters can also be created using dimensional operators. $(\phi \vee \Phi)_{1/2} : \mathcal{G} \to \mathcal{G}$ and $(\gamma \vee \Gamma)_{1/2} : \mathcal{G} \to \mathcal{G}$ filters are defined by:

$$\forall X \in \mathcal{G},\ (\phi \vee \Phi)_{1/2}(X) = (\epsilon^+_{0,1} \vee \epsilon^-_{1,0}) \circ (\delta^+_{0,1} \vee \delta^-_{1,0})(X)$$

$$\forall X \in \mathcal{G},\ (\gamma \vee \Gamma)_{1/2}(X) = (\delta^+_{0,1} \vee \delta^-_{1,0}) \circ (\epsilon^+_{0,1} \vee \epsilon^-_{1,0})(X)$$

Note that $(\phi \vee \Phi)_{1/2}$ is a closing operation, while $(\gamma \vee \Gamma)_{1/2}$ is an opening operation. These filters are illustrated in figure 14.



(a) X (b) $(\phi \vee \Phi)_{1/2}(X)$(c) $(\gamma \vee \Gamma)_{1/2}(X)$

Figure 14: Half-closing and half-opening filters.

3.2.7 Granulometries

Given parameters i and j, respectively being the quotient and the remainder of the division $k/2$, granulometric-closing filter $(\phi \vee \Phi)_{k/2} : \mathcal{G} \to \mathcal{G}$ and granulometric-opening filter $(\gamma \vee \Gamma)_{k/2} : \mathcal{G} \to \mathcal{G}$ are defined by:

$$\forall X \in \mathcal{G},\ (\phi \vee \Phi)_{k/2} = (\epsilon \vee \varepsilon)^i \circ (\phi \vee \Phi)^j_{1/2} \circ (\delta \vee \Delta)^i(X)$$

$$\forall X \in \mathcal{G},\ (\gamma \vee \Gamma)_{k/2} = (\delta \vee \Delta)^i \circ (\gamma \vee \Gamma)^j_{1/2} \circ (\epsilon \vee \varepsilon)^i(X)$$

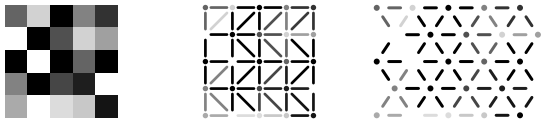
3.2.8 Alternating sequential filter

The alternating sequential filter $ASF_{k/2} : \mathcal{G} \to \mathcal{G}$ can be defined using the granulometries defined above, by:

$$\forall X \in \mathcal{G},\ ASF_{k/2}(X) = \begin{cases} \text{identity} & \text{if } k = 0 \\ (\gamma \vee \Gamma)_{k/2} \circ (\phi \vee \Phi)_{k/2} \circ ASF_{k-1/2}(X) & \text{otherwise} \end{cases}$$

3.2.9 Implementation

The graph space used in this experiment is the hexagonal grid presented in figure 11a and initialized in figure 15c, while figures 15a and 15b respectively show the image pixels and the used data structure for graph filtering. The set $\mathcal{G}_0$ is initialized with the image pixels values, while any edge $e \in \mathcal{G}_1$ is set to the minimum value of its corresponding vertices $v_1, v_2 \in \mathcal{G}_0 \mid v_1, v_2 \subset e$ with $v_1 \neq v_2$. Figure 15 illustrates the used data structure and its initialization.



(a) Image (b) Structure (c) Graph

Figure 15: Used data srtucture and initialization in graphs.

3.3 Morphological operators in vertex spaces

Dilation and erosion operators defined in vertex spaces on an hexagonal grid act as dilation and erosion operators defined in graph spaces, projected from $\mathcal{G}$ to $\mathcal{G}_0$. More precisely, the graph structure isn't implemented in these operators, as they directly act on the pixels neighbors. Besides, in this space, half-closing and half-opening operators doesn't exist and ASF filter is thus a composition of dilation and erosion operators.

3.4 Morphological area opening and closing filters

Morphological area opening and closing filters for binary and greyscale images have been presented by Vincent in [7]. These filters respectively remove light and dark regions of the image whose area is superior to a parameter $\lambda \in \mathbb{N}$. In order to compute the morphological area opening γ^a_λ (*resp.* closing ϕ^a_λ) of a greyscale image, the regional maxima (*resp.* minima) of the image are first extracted and labelled. Then, 4-neighboring pixels of each maximum (*resp.* minimum)

are progressively added relatively to their intensity, from light (*resp.* dark) to dark (*resp.* light) pixels, until the area of the broadened regions reaches λ. Pixels of the created regions are finally set to the intensity value of their last added pixel.

4 Test protocols

4.1 First test protocol

Tesseract OCR engine, presented in [6], has been used to perform optical character recognition in this experiment. This powerful system has been evaluated by UNLV-ISRI in 1995 (refer to [4]) along with other commercial OCR engines and proved its top-tier performance at the time. In order to get OCR performance results from this engine on preprocessed documents, the test data and software tools from UNLV-ISRI presented in [5] have been used. Most of the documents in this test data being available in binary format at 300 DPI resolution, the tests have been conducted on a selection of this particular set of documents (100). An other set of two generated greyscale documents has been used to evaluate the performance of each filter on greyscale values. The test procedure is basically the iteration of degradation, filtering, OCR analysis and MSE measure of each document, repeated for each couple (d, f) of degradation and filtering parameters. Note, however, that the used binary documents are scanned versions of real documents, meaning that they are imperfect and consequently contain noise. Degradation performed on these documents simply allow for a better comparison of the filters efficiency in critical conditions.

4.1.1 Degradation

Degradation levels are specified with parameter $d \in \mathbb{N}$, which acts on the binary document degradation model parameters $\Theta = (\eta, \alpha_0, \alpha, \beta_0, \beta, k)$ and the impulsional Gaussian noise parameters $\Omega = (\sigma, p)$ as follows:

$$\Theta(d) = (d * 0.02, d * 0.1, 1, d * 0.1, 1, 0)$$

$$\Omega(d) = (128, d * 0.1)$$

4.1.2 Filtering

The filters used in this experiment are the ASF filters in simplicial complexes, graphs and vertices, as well as a combination of area closing and area opening filters. The tests have been conducted on both regular and inverse versions of each document, in particular to evaluate the impact of the proposed initialization in simplicial complex spaces on binary documents. Furthermore, ASF filters in gaphs and vertices have also been evaluated with document resolution scaling of respectively 3/1 and 3/2 (Usc), in order to preserve the same number of iterations between each filter. In the case of binary document filtering, the corresponding upscaled documents were then binarized with a threshold value of 128 (Thr1), downscaled to their original size after filtering (Dsc) and binarized again with a threshold value of 128 (Thr2) in order to preserve the characters size for OCR processing. As for greyscale documents, they were also scaled but not thresholded. Filtering levels were specified for each morphological filter with parameter $f \in \mathbb{N}$. Detailed settings are described in table 1. One can note that binarization after upscaling (Thr1) is not performed in the case of vertices filtering. This is simply due to the fact that these documents are already in binary form after an exact upscaling of 3/1.

ID	Filter	Inverse	Usc	Thr1	Dsc	Thr2
1	Complex $ASF_{f/3}$					
2	Complex $ASF_{f/3}$	×				
3	Graph $ASF_{f/2}$					
4	Graph $ASF_{f/2}$	×				
5	Graph $ASF_{f/2}$		3/2	Binary docs only	2/3	Binary docs only
6	Graph $ASF_{f/2}$	×	3/2	Binary docs only	2/3	Binary docs only
7	Vertices ASF_f					
8	Vertices ASF_f	×				
9	Vertices ASF_f		3/1		1/3	Binary docs only
10	Vertices ASF_f	×	3/1		1/3	Binary docs only
11	$\phi_f^a \circ \gamma_f^a$					
12	$\phi_f^a \circ \gamma_f^a$	×				

Table 1: Filtering parameters.

4.1.3 OCR analysis

OCR analysis has been performed by the Tesseract OCR engine in its latest version (3.02). Character and word accuracy obtained from OCR processing of each document, as well as 95% confidence intervals of the obtained accuracy for each set of documents processed with a unique couple of parameters (d, f) were then computed with the accuracy, wordacc, accci and wordaccci tools provided in [5].

4.1.4 MSE measure

Mean squared error has been measured for each processed image I of dimensions $w * h$, relatively to its unprocessed counterpart O, following:

$$MSE = \frac{1}{w * h} \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} [I(i,j) - O(i,j)]^2$$

4.2 Second test protocol

Some observations can be stated about the first test protocol. One can note that MSE measures performed from binary documents that already contain noise cannot be considered as a proper evaluation of the filters denoising ability. In addition, OCR analysis and MSE measures are also affected by document scaling, which produce a slight smoothing effect that can impact the results in this situation. This second test protocol has been performed to address these two problems. However, as the characters size is a crucial factor of OCR analysis, this second test protocol is only focused on MSE performance and has thus been performed on noise-free, binary and greyscale documents that were not downscaled at all. Binary document size is 397×249 pixels while greyscale document is 481×321 pixels. The test procedure is the iteration of degradation, filtering and MSE measure of each document, repeated for each pair (d, f) of degradation and filtering parameters.

5 Results

5.1 First test protocol

In this section are shown the results of the first test protocol in the most critical tested conditions, to better compare the efficiency of each filtering setting. In figures 16, 17, 18 and 19, only the best performing setting is shown for each filter among regular and inverted document filtering.

Binary documents As can be observed in figure 16, complex filtering on non inverted documents produces better accuracy results than any other filter at original resolution. Figure 17 shows, however, that vertex filtering at triple resolution on inverted documents outperforms complex filtering for word accuracy only, but a close look at computational times reveals that this filtering method is nearly two times slower than the other. Finally, MSE results of figure 17 clearly shows that complex filtering, vertex filtering at scaled resolution and area filtering are very close, but also that graph filtering at scaled resolution is significantly outperformed on this aspect. A summary of OCR results obtained in this experiment can be observed in table 2, where ID column corresponds to the filtering parameters defined in table 1.

ID	Filter	d	f	Char acc (%)	Word acc (%)	Time (s)
1	Complexes	4	5	59.36	61.60	1354.04
2	IComplexes	4	3	13.51	32.13	879.94
3	Graphs	4	2	14.81	33.77	453.41
4	IGraphs	4	3	50.68	56.89	670.95
5	Graphs3/2	4	2	13.82	30.23	1083.59
6	IGraphs3/2	4	5	48.54	47.53	2766.04
7	Vertices	4	1	13.84	31.59	239.06
8	IVertices	4	1	40.52	50.34	342.42
9	Vertices3/1	4	2	50.79	60.72	2065.41
10	IVertices3/1	4	3	54.59	65.13	2662.26
11	Area	4	6	47.10	44.82	385.20
12	IArea	4	6	45.66	44.09	426.99

Table 2: Best OCR results of first test protocol on 100 binary documents ($d = 4$).

Greyscale documents In opposition to binary document filtering, grayscale document filtering is better performed on non inverted images for graph and vertex filtering and on inverted images for area filtering, as can be observed in figure 18. From a hierarchical point of view, complex filtering performs better than graph and vertex filtering at original resolution, with area filtering being slightly better in this conditions. Processing scaled images leads to a clear advantage of graph filtering, followed by vertex filtering, area filtering and complex filtering, but looking at computational times of vertex filtering shows that this solution has a worse performance over time ratio than complex filtering. Again, a summary of OCR results obtained in this experiment can be observed in table 3.

5.2 Second test protocol

In this section are shown the results of the second test protocol in the most critical tested conditions, to better compare the efficiency of each filtering setting. As shown in figures 20 and

ID	Filter	d	f	Char acc (%)	Word acc (%)	Time (s)
1	Complexes	4	5	90.09	67.00	41.50
2	IComplexes	4	3	68.76	43.58	26.46
3	Graphs	4	2	75.10	45.72	13.07
4	IGraphs	4	3	69.10	40.99	20.61
5	Graphs3/2	4	3	97.63	90.09	38.69
6	IGraphs3/2	4	5	82.97	60.59	77.04
7	Vertices	4	1	73.03	46.51	7.29
8	IVertices	4	2	48.23	14.30	11.52
9	Vertices3/1	4	3	90.23	78.38	74.73
10	IVertices3/1	4	4	86.42	68.13	103.85
11	Area	4	8	91.93	69.71	14.36
12	IArea	4	8	92.71	72.18	14.80

Table 3: Best OCR results of first test protocol on 2 greyscale documents ($d = 4$).

24g, complex filtering of the binary image shown in figure 24f produces better MSE results than any other filter tested in this conditions. What can also be noticed is the good performance of graph filtering at scaled resolution in these conditions, a result contrasting with the first test protocol on binary documents. Now regarding the results obtained on the greyscale image shown in figure 25f, the combination of area closing and area opening filters outperforms all other filters, as can be observed in figures 21 and 25r.

6 Discussion

It is clear that morphological operators are good tools to improve OCR accuracy performance when used as preprocessing filters. The different results shown in this experiment are potent indicators of their efficiency in the context of OCR. Indeed, preprocessing using such filters leads to an increase of respectively up to 59.34% and up to 65.12% in character and word accuracy on binary documents, while these figures are even higher in the case of greyscale documents, with respectively up to 97.63% and up to 90.09%. However, a few remarks can be stated to further explain the results obtained in this experiment. First, regarding the impact of inverted documents filtering, one can note that a general trend emerges from the results of the first test protocol. Indeed, binary document filtering is best performed in non inverted documents for complex and area filters, while graph and vertex filters performs better on inverted documents. As for greyscale document filtering, complex, graph and vertex filters are more efficient on non iverted documents, while area filter performs better on inverted documents. Second, what can also be stated is the impact of the thresholding operations done in the first test protocol on upscaled binary documents. It is clear, for instance, that graph filtering on such documents is severely impacted by these thresholds, as it is the only situation where this filter has a lower performance at a higher resolution. Third, a notable difference in the relative performance of area filtering can be observed in MSE measures of the second test protocol, when compared with complex, graph and vertex filtering. This result can be explained by the size difference of the structures in the processed documents. Indeed, as defined in section 3, morphological operators in complex, graph and vertex spaces tend to alter image contours, while it is not the case of area filters. The first group of filters is thus best performing when image contours are large and easily reconstructible, as can be observed in figure 20, while area filters are best suited for

small structures that suffer from strong contour alteration (figure 21). Finally, as a concluding remark and following the previous statement, each filter have its own strengths and weaknesses. Nevertheless, complex filtering seems to be a good choice when preprocessing binary documents with fairly large structures, while area filtering performs better on greyscale documents with small structures.

References

[1] Jean Cousty, Laurent Najman, and Jean Serra. Some morphological operators in graph spaces. In *Mathematical Morphology and Its Application to Signal and Image Processing - Proceedings of the 9th International Symposium on Mathematical Morphology (ISMM 2009)*, volume 5720 of *Lecture Notes in Computer Science*, pages 149–160, Groningen, The Netherlands, aug 2009. Springer Berlin / Heidelberg.

[2] Fábio Dias, Jean Cousty, and Laurent Najman. Some morphological operators on simplicial complex spaces. In *Discrete Geometry for Computer Imagery - Proceedings of the 16th IAPR International Conference (DGCI 2011)*, volume 6607 of *Lecture Notes in Computer Science*, pages 441–452, Nancy, France, apr 2011. Springer Berlin / Heidelberg.

[3] Tapas Kanungo, Robert M. Haralick, Henry S. Baird, Werner Stuezle, and David Madigan. A statistical, nonparametric methodology for document degradation model validation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1209–1223, nov 2000.

[4] Thomas A. Nartker, Stephen V. Rice, and Frank R. Jenkins. Ocr accuracy: Unlv's fourth annual test. *Inform*, 9(7):38–46, jul 1995.

[5] Thomas A. Nartker, Stephen V. Rice, and Steven E. Lumos. Software tools and test data for research and testing of page-reading ocr systems. In *Document Recognition and Retrieval XII*, volume 5676 of *Proceedings of SPIE*, pages 37–47, San Jose, CA, USA, jan 2005. SPIE.

[6] Ray Smith. An overview of the tesseract ocr engine. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633, Curitiba, Paraná, Brazil, sep 2007. IEEE.

[7] Luc Vincent. Morphological area openings and closings for greyscale images. In *Shape in Picture*, volume 126 of *Nato ASI Series*, pages 197–208, Driebergen, The Netherlands, sep 1992. Springer Berlin / Heidelberg.

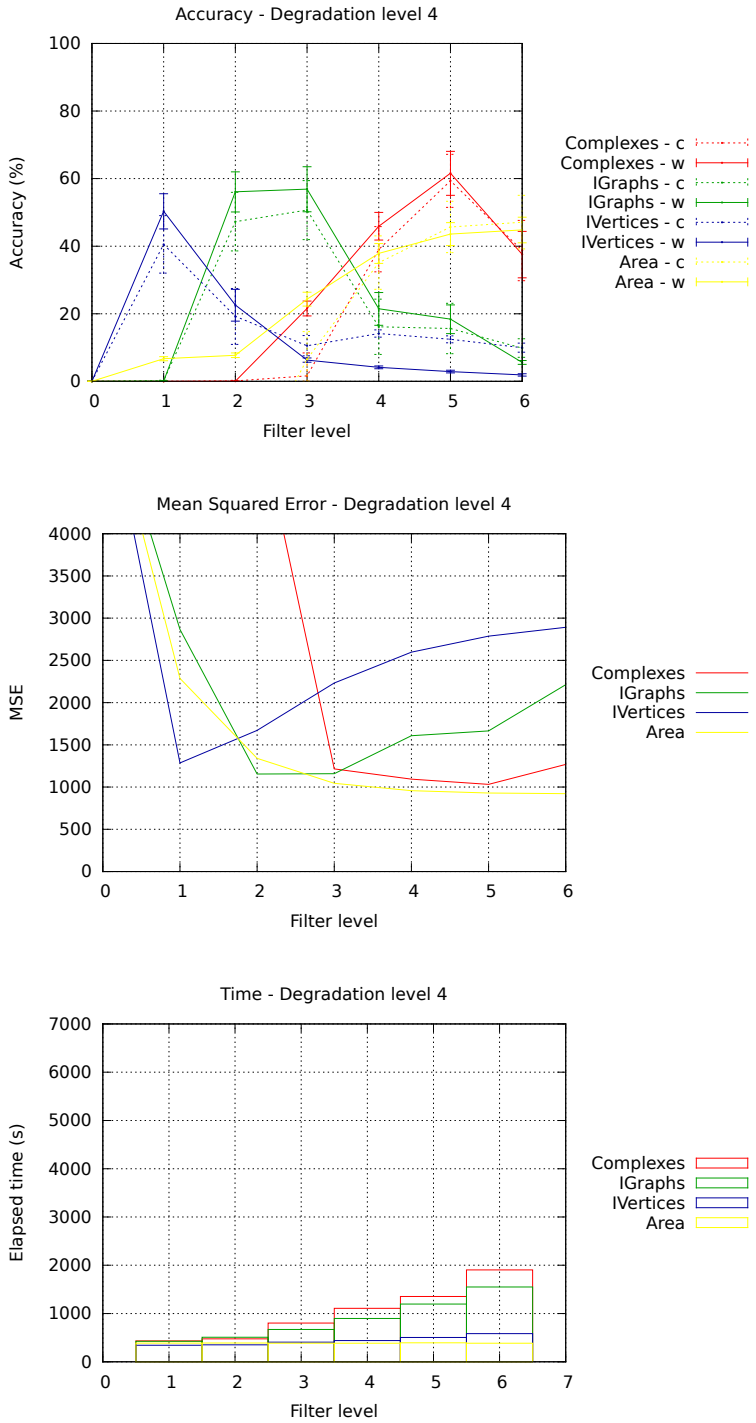


Figure 16: OCR accuracy, MSE and time measured in first test after filtering of 100 binary documents at original resolution. Solid lines represent word accuracy while dashed lines represent character accuracy.

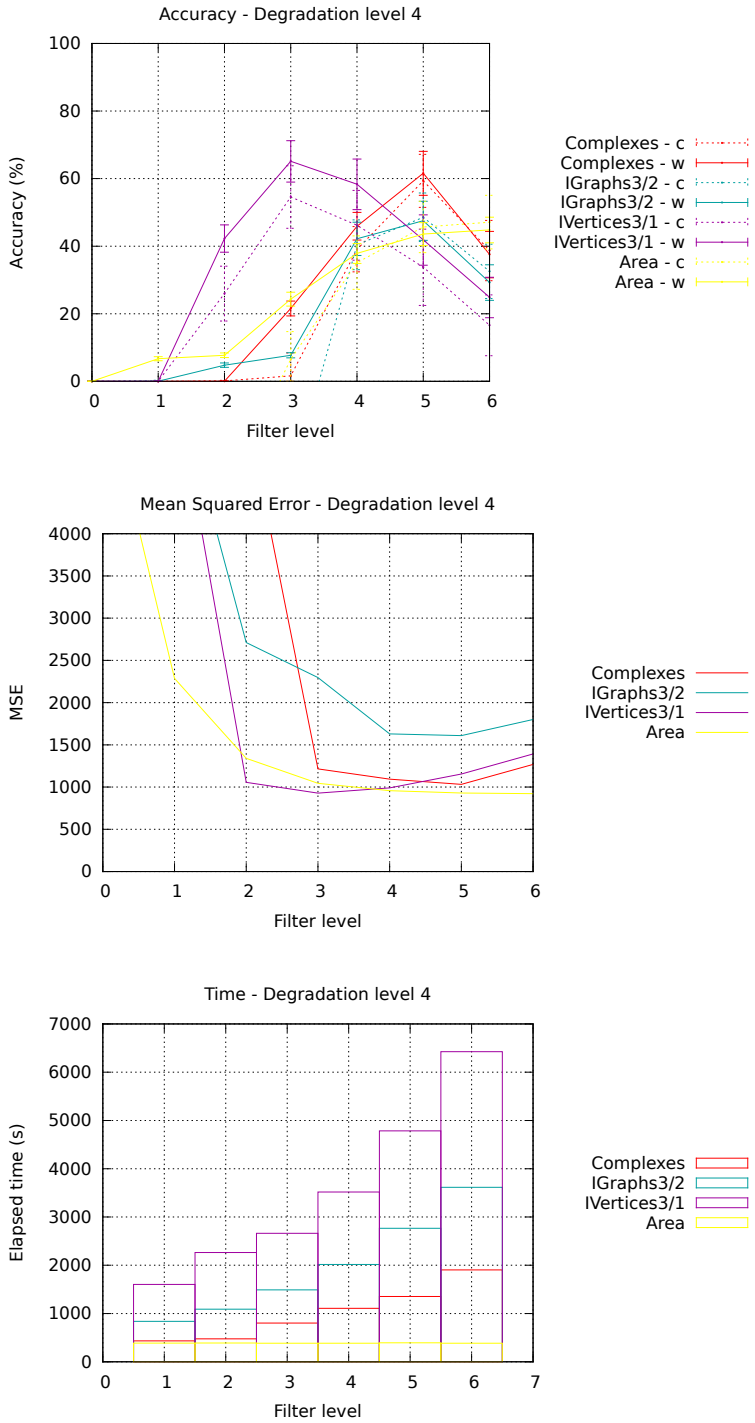


Figure 17: OCR accuracy, MSE and time measured in first test after filtering of 100 binary documents at scaled resolution for graphs and vertices filters. Solid lines represent word accuracy while dashed lines represent character accuracy.

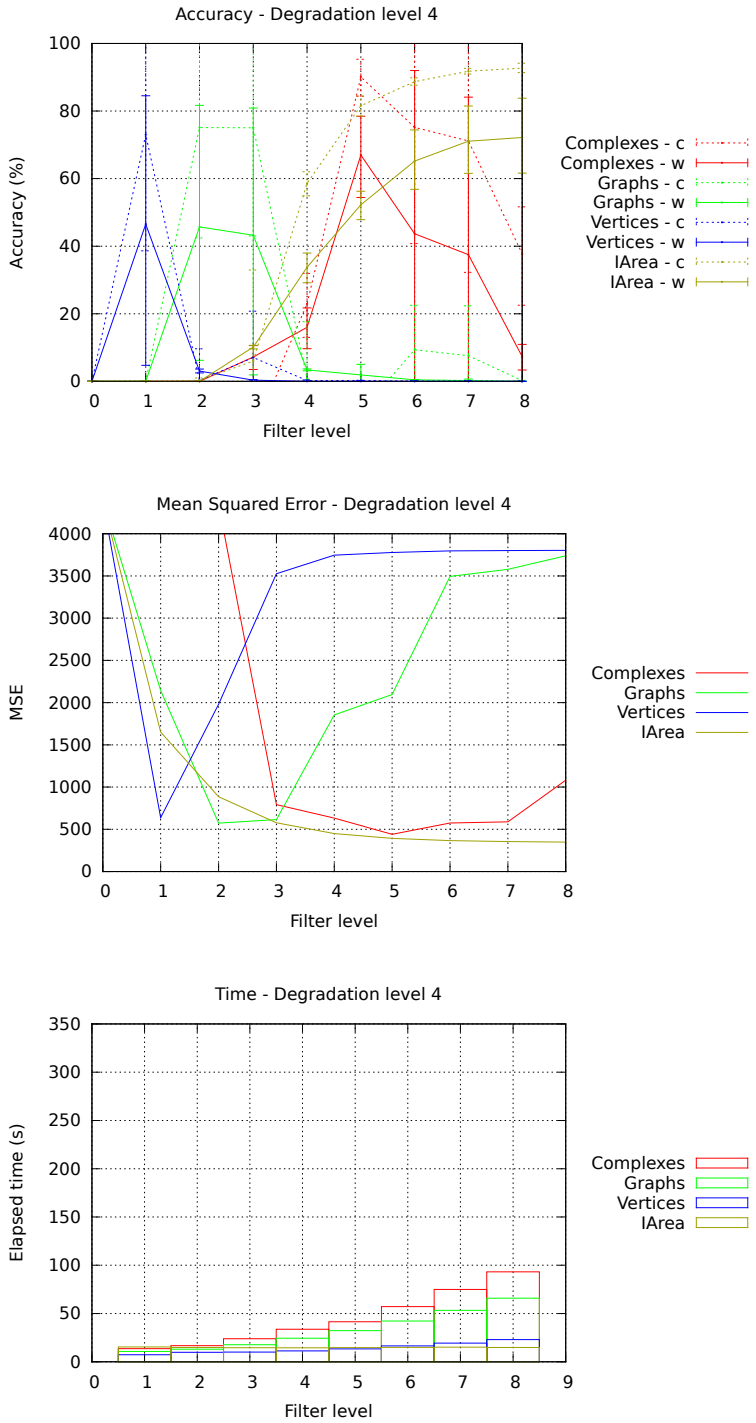


Figure 18: OCR accuracy, MSE and time measured in first test after filtering of 2 greyscale documents at original resolution. Solid lines represent word accuracy while dashed lines represent character accuracy.

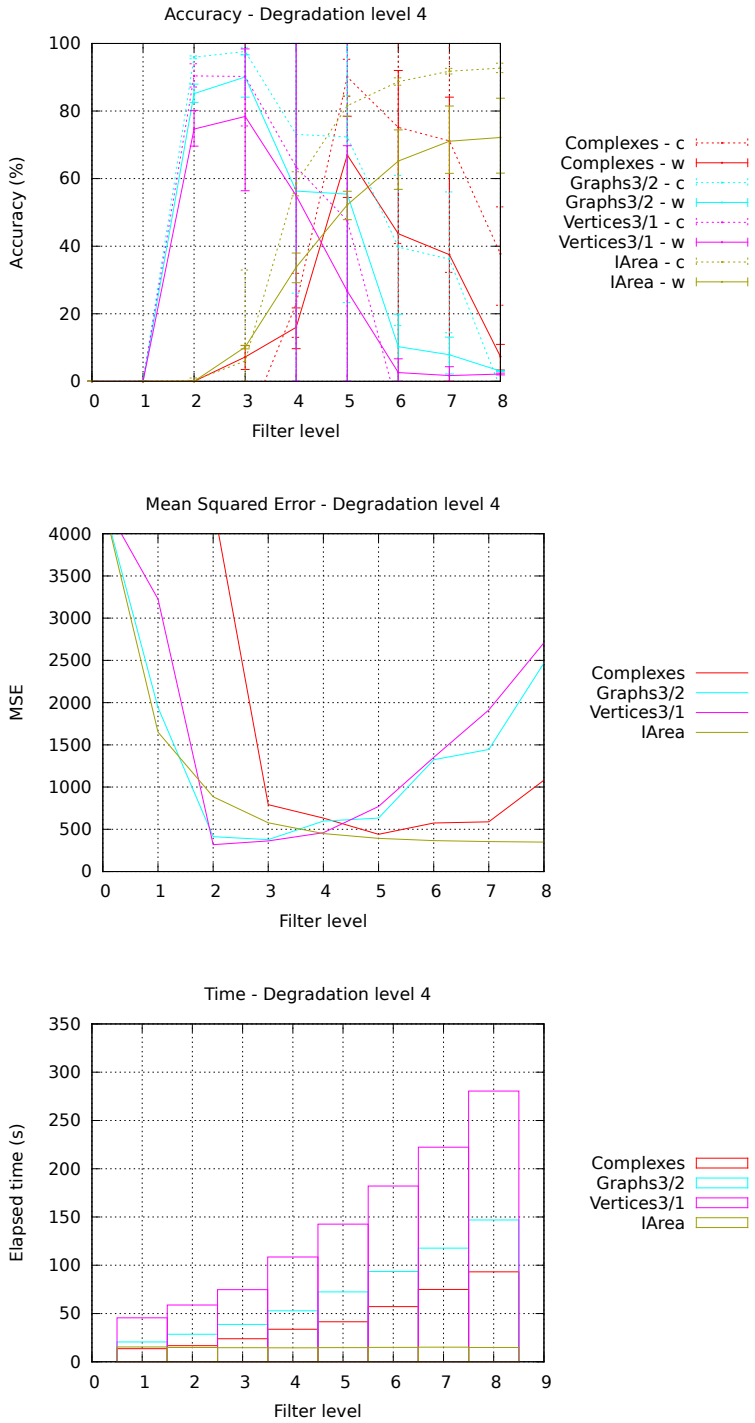


Figure 19: OCR accuracy, MSE and time measured in first test after filtering of 2 greyscale documents at scaled resolution for graphs and vertices filters. Solid lines represent word accuracy while dashed lines represent character accuracy.

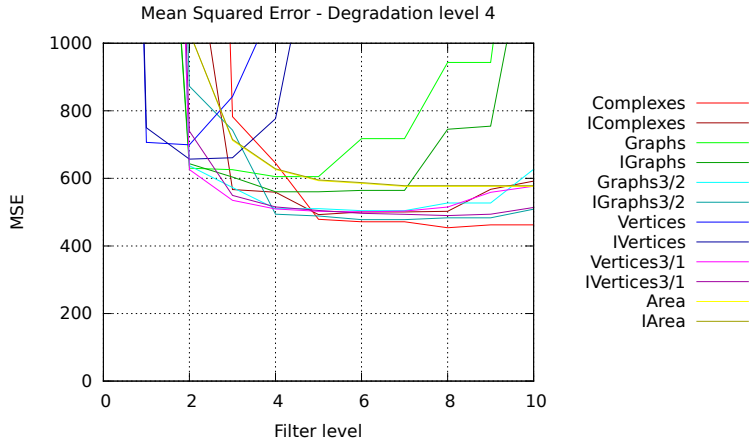


Figure 20: MSE measured in second test after filtering of the binary image shown in 24f.

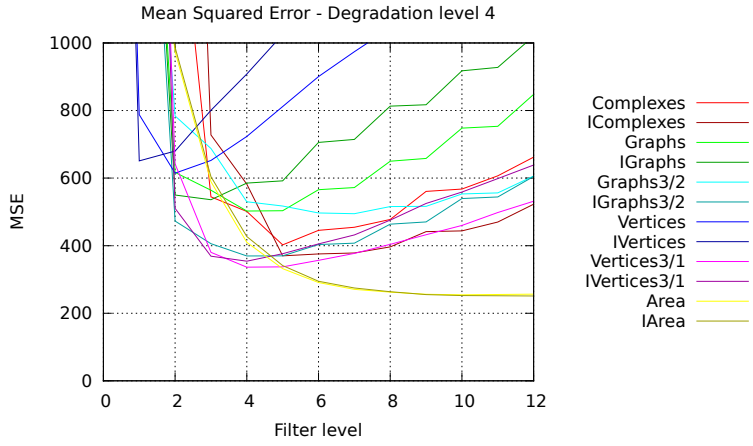
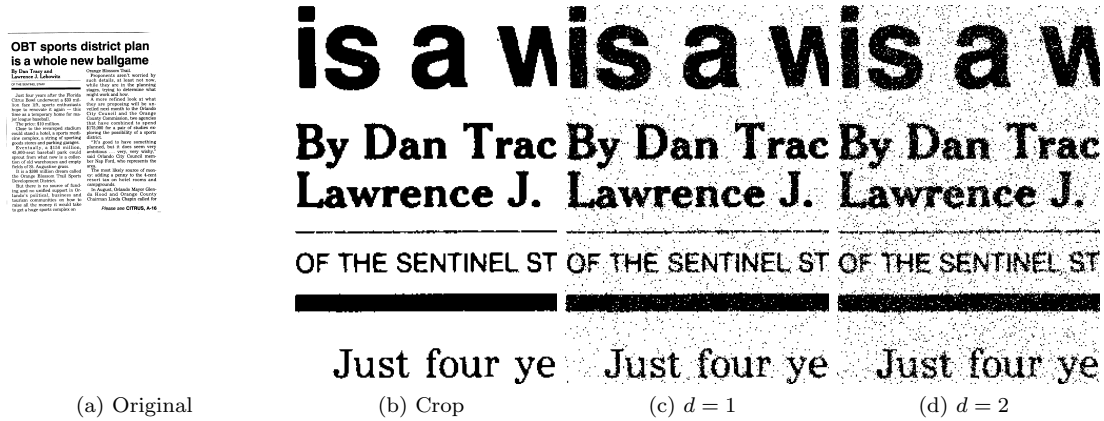
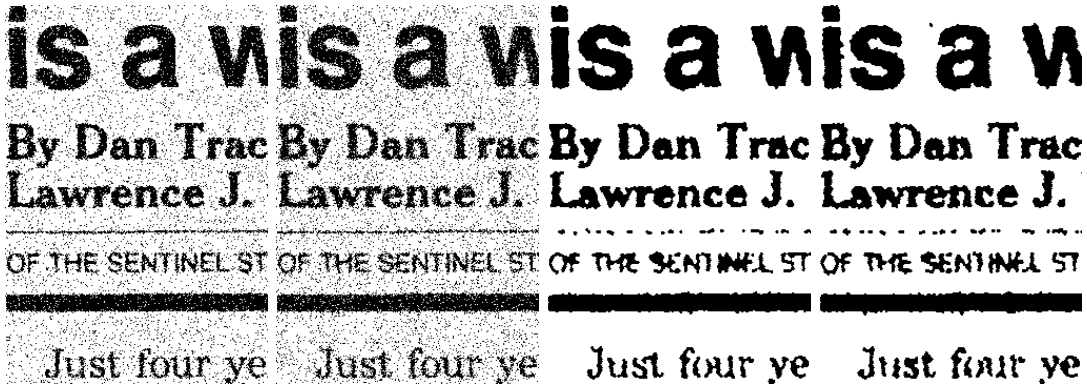


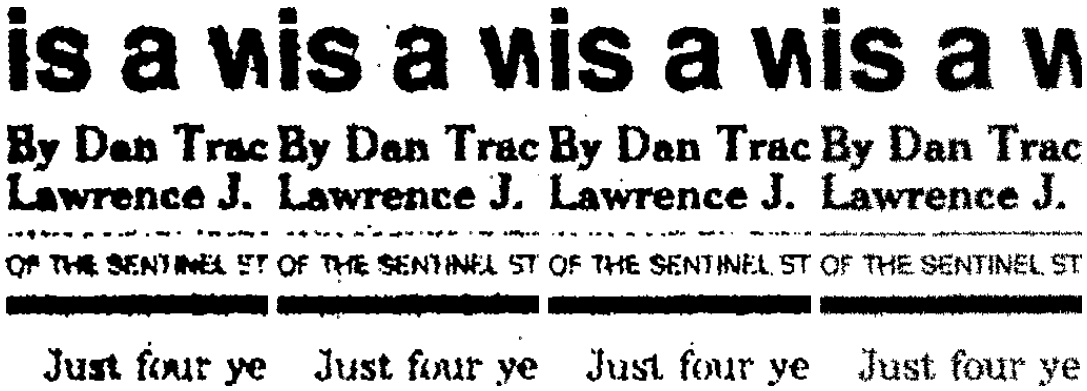
Figure 21: MSE measured in second test after filtering of the greyscale image shown in 25f.



(a) Original (b) Crop (c) $d = 1$ (d) $d = 2$

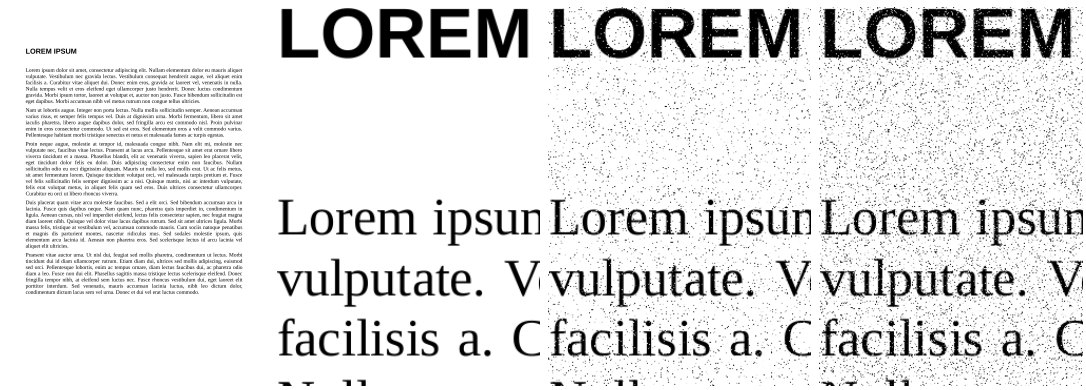


(e) $d = 3$ (f) $d = 4$ (g) $1 : 5$ (h) $4 : 3$

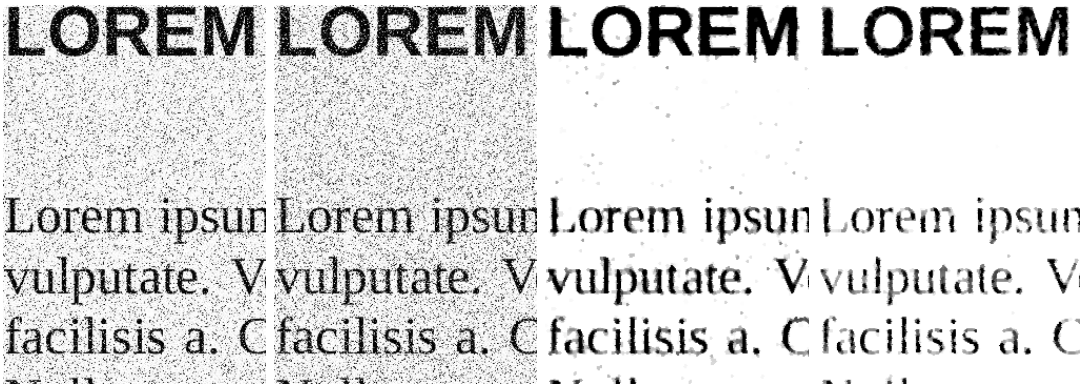


(i) $6 : 5$ (j) $8 : 1$ (k) $10 : 3$ (l) $11 : 6$

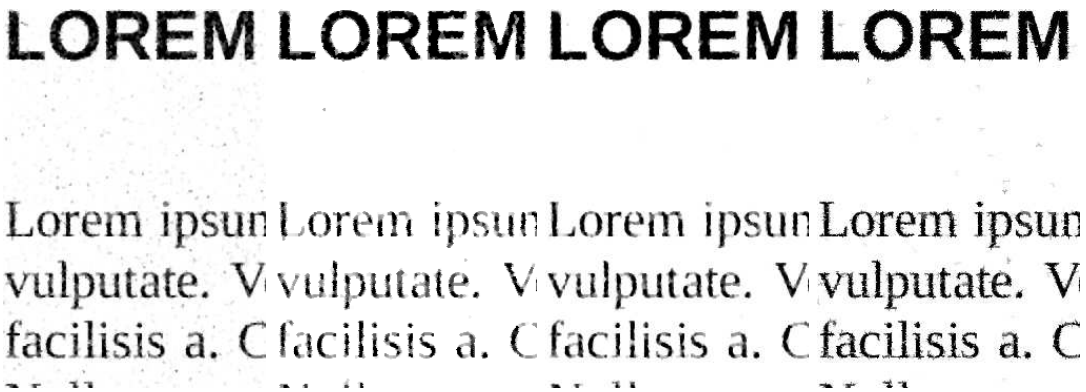
Figure 22: First test protocol sample on binary documents. Original and degraded images with binary degradation model, along with best filtering results obtained on image 22f, under the form [ID : f].



(a) Original (b) Crop (c) $d = 1$ (d) $d = 2$

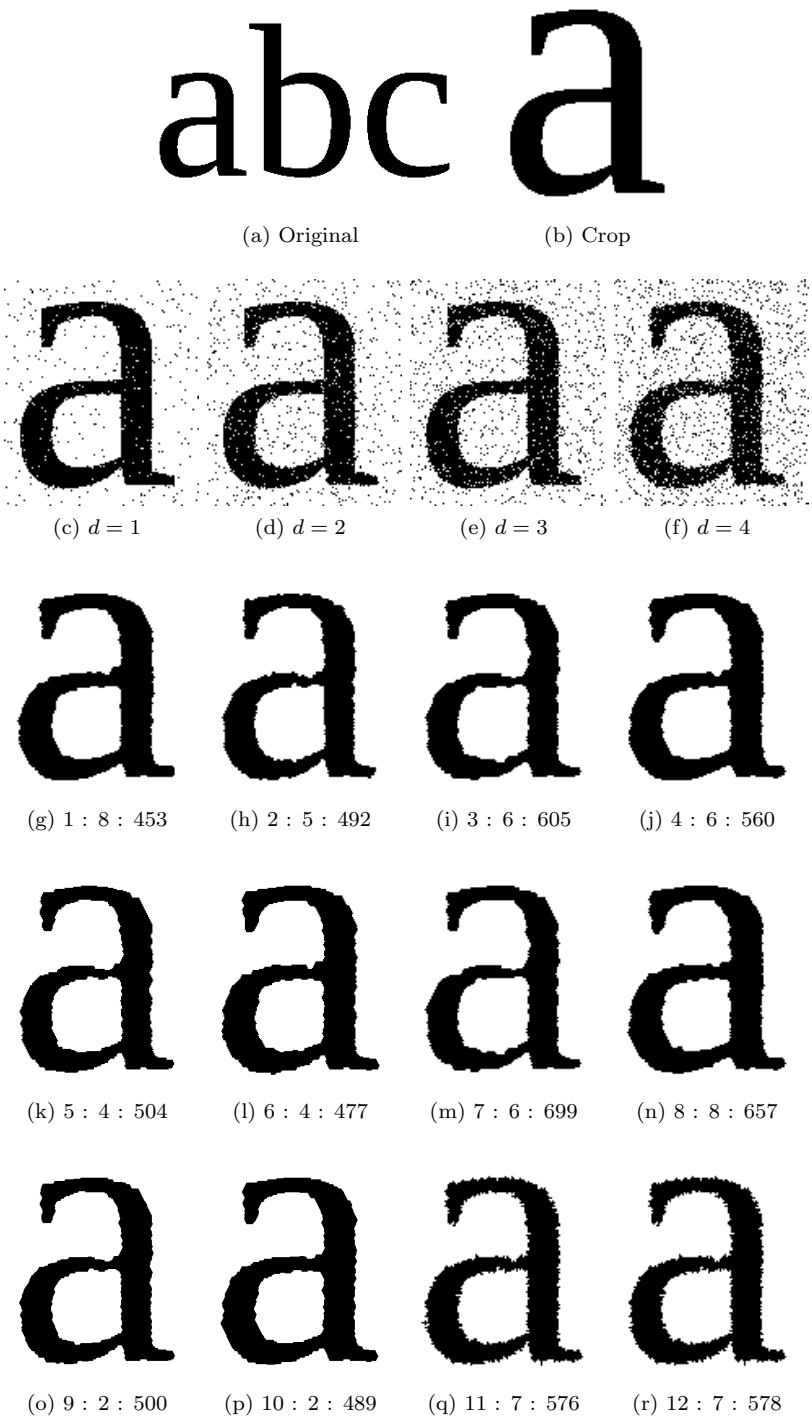


(e) $d = 3$ (f) $d = 4$ (g) $1 : 5$ (h) $3 : 2$



(i) $5 : 3$ (j) $7 : 1$ (k) $9 : 3$ (l) $12 : 8$

Figure 23: First test protocol sample on greyscale documents. Original and degraded images with impulsional Gaussian noise, along with best filtering results obtained on image 23f, under the form [ID : f].



(a) Original (b) Crop

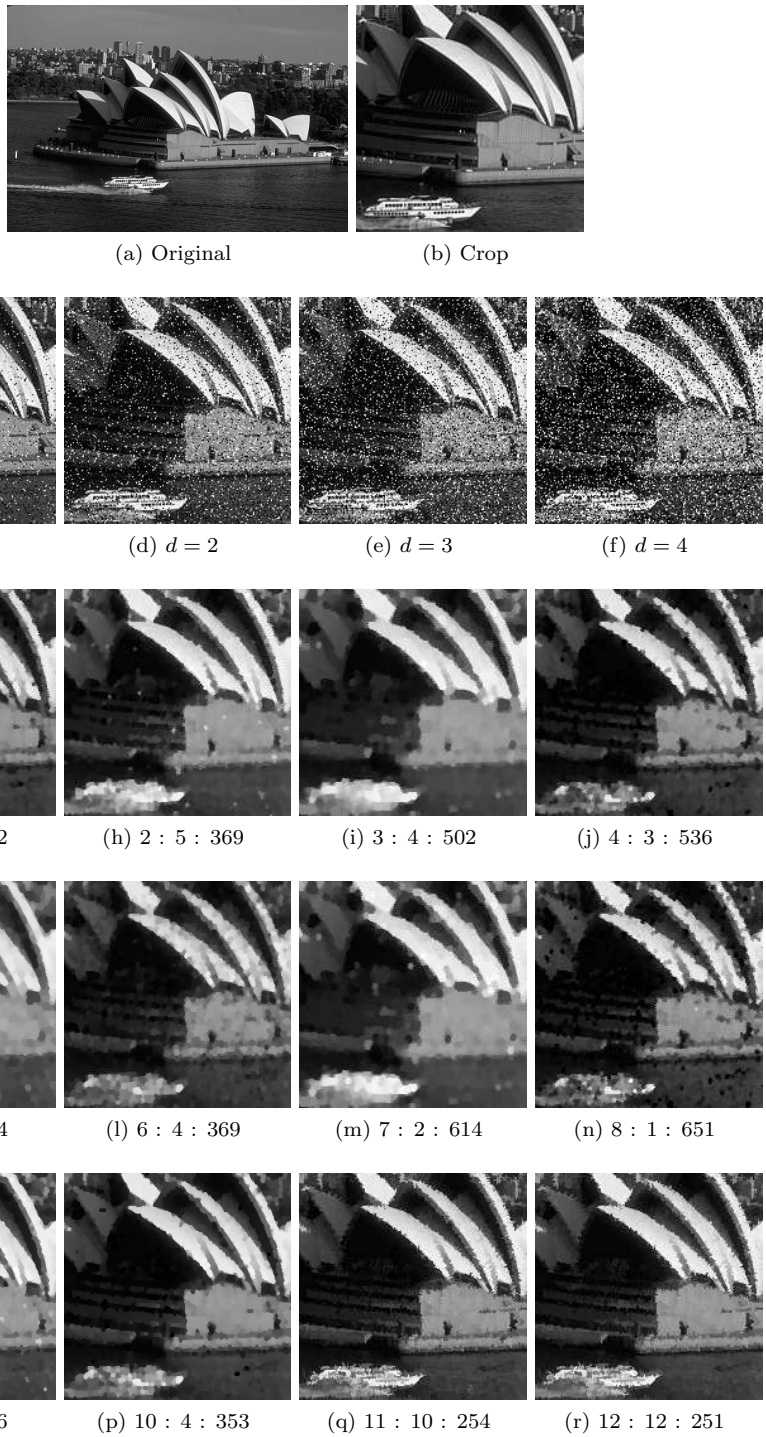
(c) $d = 1$ (d) $d = 2$ (e) $d = 3$ (f) $d = 4$

(g) 1 : 8 : 453 (h) 2 : 5 : 492 (i) 3 : 6 : 605 (j) 4 : 6 : 560

(k) 5 : 4 : 504 (l) 6 : 4 : 477 (m) 7 : 6 : 699 (n) 8 : 8 : 657

(o) 9 : 2 : 500 (p) 10 : 2 : 489 (q) 11 : 7 : 576 (r) 12 : 7 : 578

Figure 24: Second test protocol sample on binary document. Original and degraded images with binary degradation model, along with best filtering results obtained on image 24f ($d = 4$), under the form [ID : f : MSE].



(a) Original (b) Crop

(c) $d = 1$ (d) $d = 2$ (e) $d = 3$ (f) $d = 4$

(g) 1 : 5 : 402 (h) 2 : 5 : 369 (i) 3 : 4 : 502 (j) 4 : 3 : 536

(k) 5 : 7 : 494 (l) 6 : 4 : 369 (m) 7 : 2 : 614 (n) 8 : 1 : 651

(o) 9 : 4 : 336 (p) 10 : 4 : 353 (q) 11 : 10 : 254 (r) 12 : 12 : 251

Figure 25: Second test protocol sample on greyscale document. Original and degraded images with impulsional Gaussian noise, along with best filtering results obtained on image 25f ($d = 4$), under the form [ID : f : MSE].